Linear Regression

Nipun Batra and the teaching staff

IIT Gandhinagar

August 1, 2025

Table of Contents

Linear Regression

- Output is continuous in nature.

Linear Regression

- Output is continuous in nature.
- Examples of linear systems:

Linear Regression

- Output is continuous in nature.
- Examples of linear systems:
  - $F = ma$

Linear Regression

- Output is continuous in nature.
- Examples of linear systems:
  - $F = ma$
  - $v = u + at$

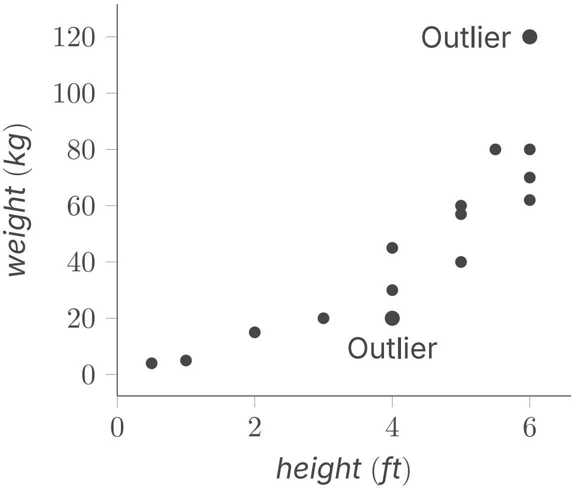
Task at hand

- TASK: Predict Weight = f(height)

Height	Weight
3	29
4	35
5	39
2	20
6	41
7	?
8	?
1	?

The first part of the dataset is the training points. The latter ones are testing points.

Scatter Plot



Matrix representation of the expression

- $\textit{weight}_1 \approx \theta_0 + \theta_1 \cdot \textit{height}_1$

Matrix representation of the expression

- $\textit{weight}_1 \approx \theta_0 + \theta_1 \cdot \textit{height}_1$
- $\textit{weight}_2 \approx \theta_0 + \theta_1 \cdot \textit{height}_2$

Matrix representation of the expression

- $\textit{weight}_1 \approx \theta_0 + \theta_1 \cdot \textit{height}_1$
- $\textit{weight}_2 \approx \theta_0 + \theta_1 \cdot \textit{height}_2$
- $\textit{weight}_N \approx \theta_0 + \theta_1 \cdot \textit{height}_N$

Matrix representation of the expression

- $\textit{weight}_1 \approx \theta_0 + \theta_1 \cdot \textit{height}_1$
- $\textit{weight}_2 \approx \theta_0 + \theta_1 \cdot \textit{height}_2$
- $\textit{weight}_N \approx \theta_0 + \theta_1 \cdot \textit{height}_N$

Matrix representation of the expression

- $\mathit{weight}_1 \approx \theta_0 + \theta_1 \cdot \mathit{height}_1$
- $\mathit{weight}_2 \approx \theta_0 + \theta_1 \cdot \mathit{height}_2$
- $\mathit{weight}_N \approx \theta_0 + \theta_1 \cdot \mathit{height}_N$

$$\boxed{\mathit{weight}_i \approx \theta_0 + \theta_1 \cdot \mathit{height}_i}$$

Matrix representation of the expression

$$\begin{bmatrix} \textit{weight}_1 \\ \textit{weight}_2 \\ \dots \\ \textit{weight}_N \end{bmatrix} = \begin{bmatrix} 1 & \textit{height}_1 \\ 1 & \textit{height}_2 \\ \dots & \dots \\ 1 & \textit{height}_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n \times 1} = \mathbf{X}_{n \times d} \boldsymbol{\theta}_{d \times 1}$$

Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times d}\boldsymbol{\theta}_{d\times 1}$$

- θ_0 - Bias Term/Intercept Term

Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n\times 1} = \mathbf{X}_{n\times d}\boldsymbol{\theta}_{d\times 1}$$

- θ_0 - Bias Term/Intercept Term
- θ_1 - Slope

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.
Examples in multiple dimensions.

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.
Examples in multiple dimensions.
One example is to predict the water demand of the IITGN campus

Extension to multiple dimensions

In the previous example y = f(x), where x is one-dimensional.
Examples in multiple dimensions.
One example is to predict the water demand of the IITGN campus

Demand = f(# occupants, Temperature)

Extension to multiple dimensions

In the previous example y = f(x), where x is one-dimensional.
Examples in multiple dimensions.
One example is to predict the water demand of the IITGN campus

Demand = f(# occupants, Temperature)

Demand = Base Demand + K_1 * # occupants + K_2 * Temperature

Intuition

We hope to:

- Learn f: *Demand* = $f(\#occupants, Temperature)$

Intuition

We hope to:

- Learn f: $Demand = f(\#occupants, Temperature)$
- From training dataset

Intuition

We hope to:

- Learn *f*: *Demand* = $f(\#occupants, Temperature)$
- From training dataset
- To predict the condition for the testing set

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is $\hat{demand}_i = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is $\hat{demand}_i = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$
- $\hat{demand}_i = x_i'^T \theta$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is $\hat{demand}_i = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$
- $\hat{demand}_i = x_i'^T \theta$
- where $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is $\hat{demand_i} = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$
- $\hat{demand_i} = x_i'^T \theta$
- where $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$
- and $x_i' = \begin{bmatrix} 1 \\ Temperature_i \\ \#Occupants_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is $\hat{demand}_i = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$
- $\hat{demand}_i = x_i'^T \theta$
- where $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$
- and $x_i' = \begin{bmatrix} 1 \\ Temperature_i \\ \#Occupants_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$
- Notice the transpose in the equation! This is because x_i is a column vector

We can expect the following

- Demand increases, if # occupants increases, then θ_2 is likely to be positive

We can expect the following

- Demand increases, if # occupants increases, then θ_2 is likely to be positive
- Demand increases, if temperature increases, then θ_1 is likely to be positive

We can expect the following

- Demand increases, if # occupants increases, then θ_2 is likely to be positive
- Demand increases, if temperature increases, then θ_1 is likely to be positive
- Base demand is independent of the temperature and the # occupants, but, likely positive, thus θ_0 is likely positive.

Generalized Linear Regression Format

- Assuming N samples for training

Generalized Linear Regression Format

- Assuming N samples for training
- # Features = M

Generalized Linear Regression Format

- Assuming N samples for training
- # Features = M

Generalized Linear Regression Format

- Assuming N samples for training
- # Features = M

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N\times(M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1)\times 1}$$

Generalized Linear Regression Format

- Assuming N samples for training
- # Features = M

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N\times(M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1)\times 1}$$

$$\boxed{\hat{Y} = X\theta}$$

Relationships between feature and target variables

- There could be different $\theta_0, \theta_1 \ldots \theta_M$. Each of them can represents a relationship.

Relationships between feature and target variables

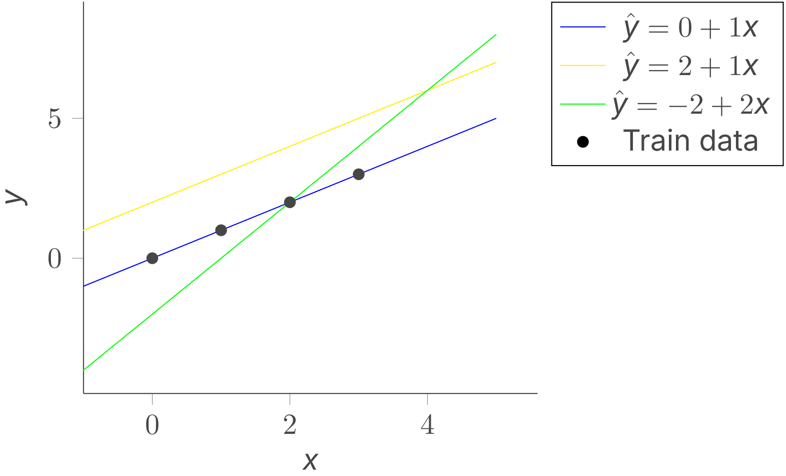
- There could be different $\theta_0, \theta_1 \dots \theta_M$. Each of them can represents a relationship.
- Given multiples values of $\theta_0, \theta_1 \dots \theta_M$ how to choose which is the best?

Relationships between feature and target variables

- There could be different $\theta_0, \theta_1 \dots \theta_M$. Each of them can represents a relationship.
- Given multiples values of $\theta_0, \theta_1 \dots \theta_M$ how to choose which is the best?
- Let us consider an example in 2d

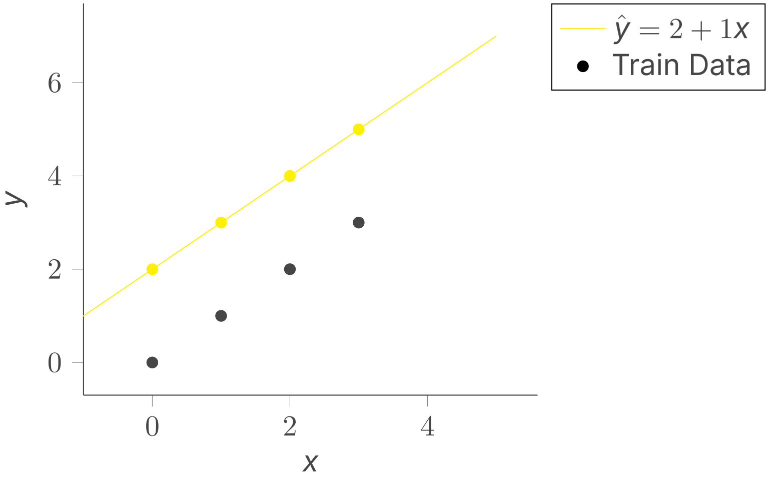
Relationships between feature and target variables

Out of the three fits, which one do we choose?



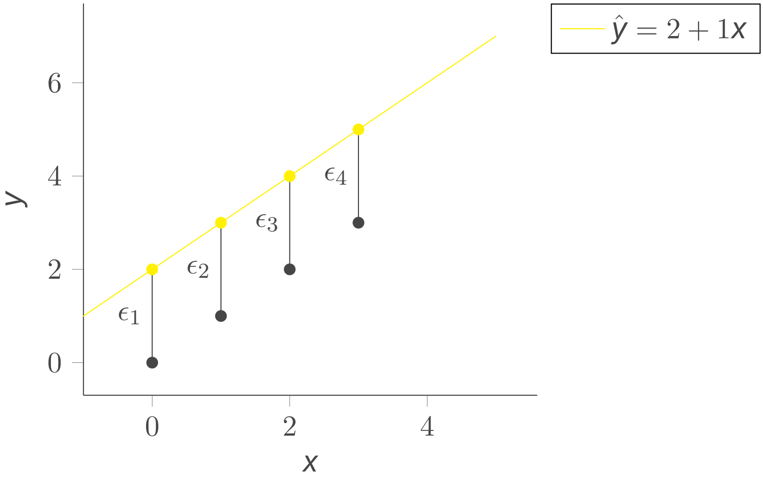
Relationships between feature and target variables

We have $\hat{y} = 2 + 1x$ as one relationship.



Relationships between feature and target variables

How far is our estimated $\hat{y}$ from ground truth y?



Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)
- y_i denotes the ground truth for i^{th} sample

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = \mathrm{x}_i^\top \boldsymbol{\theta}$
- ϵ_i denotes the error/residual for i^{th} sample

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = \mathrm{x}_i^\top \boldsymbol{\theta}$
- ϵ_i denotes the error/residual for i^{th} sample
- θ_0, θ_1: The parameters of the linear regression

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- ϵ_i denotes the error/residual for i^{th} sample
- θ_0, θ_1: The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption**: ϵ_i are independent and identically distributed (i.i.d.)
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- ϵ_i denotes the error/residual for i^{th} sample
- θ_0, θ_1: The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$
- $\epsilon_i = y_i - (\theta_0 + x_i \cdot \theta_1)$

Good fit

- $|\epsilon_1|$, $|\epsilon_2|$, $|\epsilon_3|$, ... should be small.

Good fit

- $|\epsilon_1|$, $|\epsilon_2|$, $|\epsilon_3|$, ... should be small.
- minimize $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$ - L_2 Norm

Good fit

- $|\epsilon_1|$, $|\epsilon_2|$, $|\epsilon_3|$, ... should be small.
- minimize $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$ - L_2 Norm
- minimize $|\epsilon_1| + |\epsilon_2| + \cdots + |\epsilon_n|$ - L_1 Norm

Normal Equation

Normal Equation

$$Y = X\theta + \epsilon$$

Normal Equation

$$Y = X\theta + \epsilon$$

To Learn: θ

Normal Equation

$$Y = X\theta + \epsilon$$

To Learn: θ
Objective: minimize $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$

Normal Equation

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

Normal Equation

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

Objective: Minimize $\epsilon^T \epsilon$

Derivation of Normal Equation

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$$

$$\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$$

This is what we wish to minimize

Minimizing the objective function

$$\frac{\partial \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}} = 0$$

- $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{y}^\top \mathbf{y} = 0$

Substitute the values in the top equation

Minimizing the objective function

$$\frac{\partial \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}} = 0$$

- $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{y}^\top \mathbf{y} = 0$
- $\frac{\partial}{\partial \boldsymbol{\theta}} (-2\mathbf{y}^\top \mathbf{X} \boldsymbol{\theta}) = -2\mathbf{X}^\top \mathbf{y}$

Substitute the values in the top equation

Minimizing the objective function

$$\frac{\partial \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}} = 0$$

- $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{y}^\top \mathbf{y} = 0$
- $\frac{\partial}{\partial \boldsymbol{\theta}} (-2\mathbf{y}^\top \mathbf{X} \boldsymbol{\theta}) = -2\mathbf{X}^\top \mathbf{y}$
- $\frac{\partial}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}) = 2\mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}$

Substitute the values in the top equation

Normal Equation derivation

$$0 = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$$

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$$

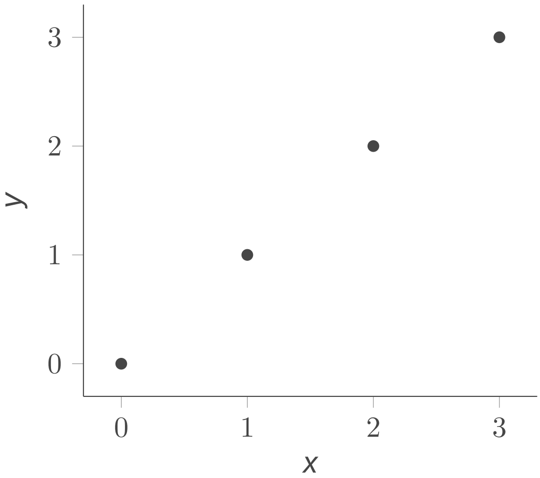
$$\boxed{\hat{\boldsymbol{\theta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}}$$

Worked out example

x	y
0	0
1	1
2	2
3	3

Given the data above, find θ_0 and θ_1.

Scatter Plot



Worked out example

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$$

Given the data above, find θ_0 and θ_1.

Worked out example

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}$$

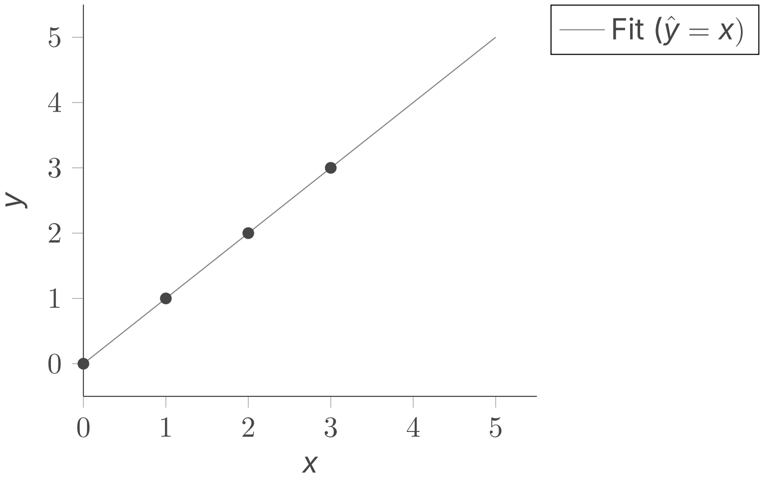
$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

Worked out example

$$\theta = (X^T X)^{-1}(X^T y)$$

$$\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix} = \frac{1}{20}\begin{bmatrix}14 & -6\\ -6 & 4\end{bmatrix}\begin{bmatrix}6\\ 14\end{bmatrix} = \begin{bmatrix}0\\ 1\end{bmatrix}$$

Scatter Plot

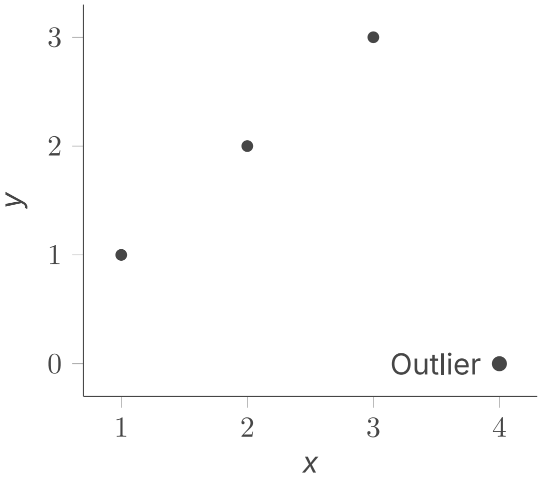


Effect of outlier

x	y
1	1
2	2
3	3
4	0

Compute the θ_0 and θ_1.

Scatter Plot



Worked out example

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

Given the data above, find θ_0 and θ_1.

Worked out example

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

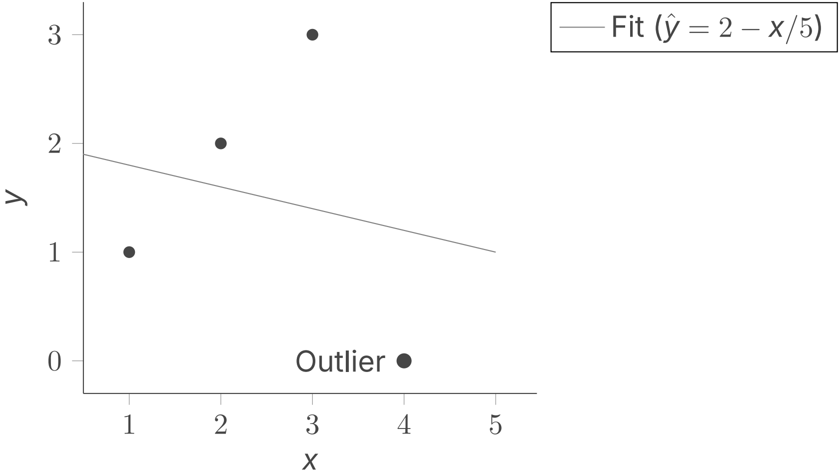
$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

Worked out example

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{y})$$

$$\begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix} = \begin{bmatrix}2\\ (-1/5)\end{bmatrix}$$

Scatter Plot



Variable Transformation

Transform the data, by including the higher power terms in the feature space.

t	s
0	0
1	6
3	24
4	36

The above table represents the data before transformation

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation
Now, we can write $\hat{s} = f(t, t^2)$

Variable Transformation

Add the higher degree features to the previous table

t	t^2	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Now, we can write $\hat{s} = f(t, t^2)$

Other transformations: $\log(x), x_1 \times x_2$

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?
4. Is $\hat{s} = \theta_0 + \theta_1 * t + \mathrm{e}^{\theta_2} * t$ linear?

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?
4. Is $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$ linear?
5. All except #4 are linear models!

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?
3. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$ linear?
4. Is $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$ linear?
5. All except #4 are linear models!
6. Linear refers to the relationship between the parameters that you are estimating (θ) and the outcome

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

Basis Functions

- Linear regression only refers to linear in the parameters

Basis Functions

- Linear regression only refers to linear in the parameters
- We can perform an arbitrary nonlinear transformation $\phi(x)$ of the inputs x and then linearly combine the components of this transformation.

Basis Functions

- Linear regression only refers to linear in the parameters
- We can perform an arbitrary nonlinear transformation $\phi(x)$ of the inputs x and then linearly combine the components of this transformation.
- $\phi : \mathbb{R}^D \to \mathbb{R}^K$ is called the basis function

Basis Functions

Some examples of basis functions:

- Polynomial basis: $\phi(x) = \{1, x, x^2, x^3, \dots\}$

Basis Functions

Some examples of basis functions:

- Polynomial basis: $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:
  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$

Basis Functions

Some examples of basis functions:

- Polynomial basis: $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:
  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$
- Gaussian basis:
  $\phi(x) = \{1, \exp(-\frac{(x-\mu_1)^2}{2\sigma^2}), \exp(-\frac{(x-\mu_2)^2}{2\sigma^2}), \dots\}$

Basis Functions

Some examples of basis functions:

- Polynomial basis: $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:
  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$
- Gaussian basis:
  $\phi(x) = \{1, \exp(-\frac{(x-\mu_1)^2}{2\sigma^2}), \exp(-\frac{(x-\mu_2)^2}{2\sigma^2}), \dots\}$
- Sigmoid basis: $\phi(x) = \{1, \sigma(x - \mu_1), \sigma(x - \mu_2), \dots\}$
  where $\sigma(x) = \frac{1}{1+e^{-x}}$

Linear Combination of Vectors

Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ be vectors in $\mathbb{R}^D$, where D denotes the dimensions.

Linear Combination of Vectors

Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ be vectors in $\mathbb{R}^D$, where D denotes the dimensions.
A linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ is of the following form

Linear Combination of Vectors

Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ be vectors in $\mathbb{R}^D$, where D denotes the dimensions.
A linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_i$ is of the following form

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \alpha_3 \mathbf{v}_3 + \cdots + \alpha_i \mathbf{v}_i$$

where $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_i \in \mathbb{R}$

Span of vectors

Let $v_1, v_2, \ldots, v_i$ be vectors in $\mathbb{R}^D$, with D dimensions.

Span of vectors

Let $v_1, v_2, \ldots, v_i$ be vectors in $\mathbb{R}^D$, with D dimensions.
The span of $v_1, v_2, \ldots, v_i$ is denoted by
$\text{SPAN}\{v_1, v_2, \ldots, v_i\}$

Span of vectors

Let $v_1, v_2, \ldots, v_i$ be vectors in $\mathbb{R}^D$, with D dimensions.
The span of $v_1, v_2, \ldots, v_i$ is denoted by
SPAN$\{v_1, v_2, \ldots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

Span of vectors

Let $v_1, v_2, \ldots, v_i$ be vectors in $\mathbb{R}^D$, with D dimensions.
The span of $v_1, v_2, \ldots, v_i$ is denoted by
$\text{SPAN}\{v_1, v_2, \ldots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

It is the set of all vectors that can be generated by linear combinations of $v_1, v_2, \ldots, v_i$.

Span of vectors

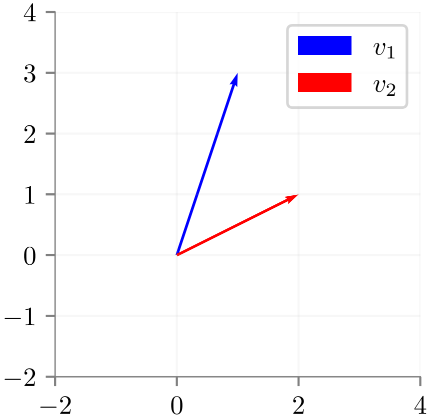
Let $v_1, v_2, \ldots, v_i$ be vectors in $\mathbb{R}^D$, with D dimensions.
The span of $v_1, v_2, \ldots, v_i$ is denoted by
SPAN$\{v_1, v_2, \ldots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

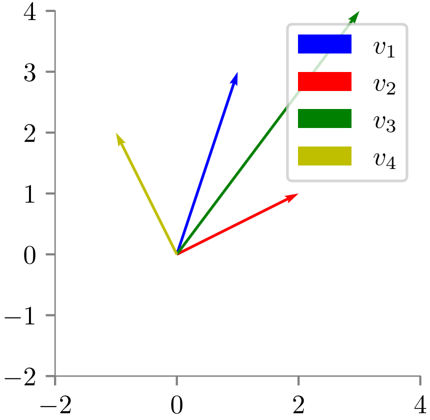
It is the set of all vectors that can be generated by linear combinations of $v_1, v_2, \ldots, v_i$.
If we stack the vectors $v_1, v_2, \ldots, v_i$ as columns of a matrix V, then the span of $v_1, v_2, \ldots, v_i$ is given as $V\alpha$ where $\alpha \in \mathbb{R}^i$

Example

Find the span of $\left(\begin{bmatrix}1\\3\end{bmatrix}, \begin{bmatrix}2\\1\end{bmatrix}\right)$



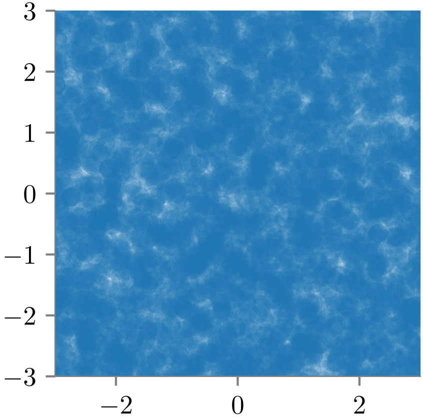
Example



We have $v_3 = v_1 + v_2$
We have $v_4 = v_1 - v_2$

Example

Simulating the above example in python using different values of α_1 and α_2



Span$((v_1, v_2)) \in \mathcal{R}^2$

Example

Find the span of $\left(\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2\\4\end{bmatrix}\right)$

Example

Find the span of ($\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix}$)

Can we obtain a point (x, y) s.t. x = 3y?

Example

Find the span of ($\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix})$

Can we obtain a point (x, y) s.t. x = 3y?

No

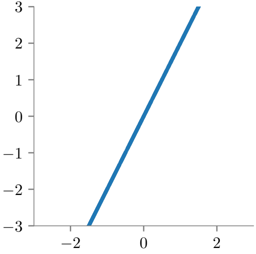
Example

Find the span of ($\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix}$)

Can we obtain a point (x, y) s.t. x = 3y?

No

Span of the above set is along the line y = 2x

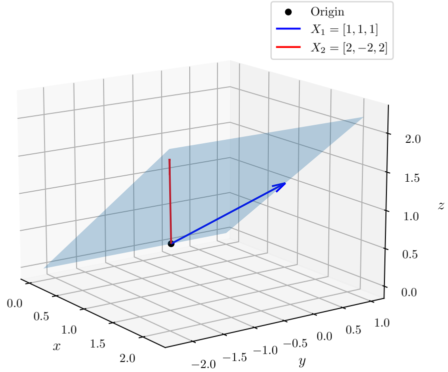


Example

Find the span of $\left(\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}2\\-2\\2\end{bmatrix}\right)$

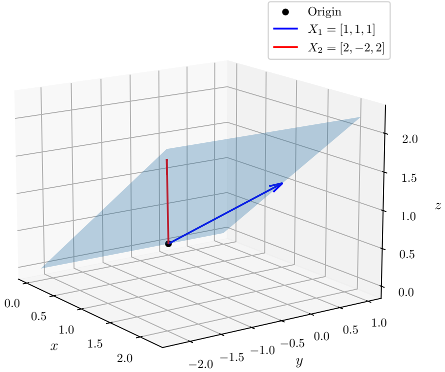
Example

Find the span of ($\begin{bmatrix}1\\1\\1\end{bmatrix}$, $\begin{bmatrix}2\\-2\\2\end{bmatrix}$)



Example

Find the span of $\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$



The span is the plane $z = x$ or $x_3 = x_1$

Geometric Interpretation

Consider $\mathbf{X}$ and $\mathbf{y}$ as follows.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

- We are trying to learn $\boldsymbol{\theta}$ for $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$ such that $||\mathbf{y} - \hat{\mathbf{y}}||_2$ is minimised

Geometric Interpretation

Consider $\mathbf{X}$ and $\mathbf{y}$ as follows.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

- We are trying to learn $\boldsymbol{\theta}$ for $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$ such that $||\mathbf{y} - \hat{\mathbf{y}}||_2$ is minimised
- Consider the two columns of $\mathbf{X}$. Can we write $\mathbf{X}\boldsymbol{\theta}$ as the span of ($\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$)?

Geometric Interpretation

Consider $\mathbf{X}$ and $\mathbf{y}$ as follows.

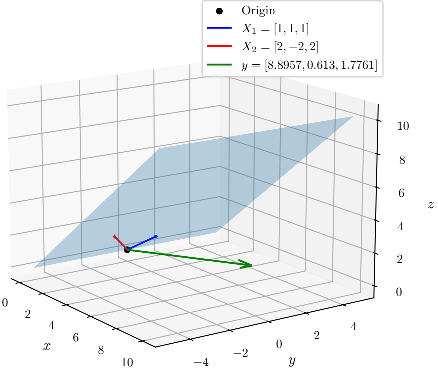
$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

- We are trying to learn $\boldsymbol{\theta}$ for $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$ such that $||\mathbf{y} - \hat{\mathbf{y}}||_2$ is minimised
- Consider the two columns of $\mathbf{X}$. Can we write $\mathbf{X}\boldsymbol{\theta}$ as the span of ($\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$)?
- We wish to find $\hat{\mathbf{y}}$ such that

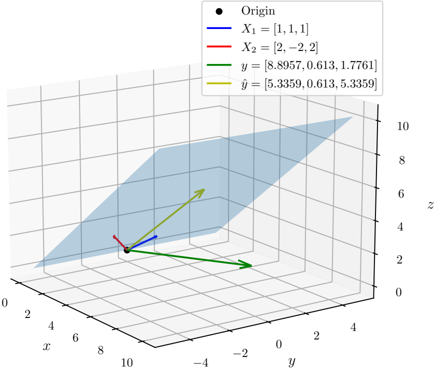
$$\underset{\hat{\mathbf{y}} \in SPAN\{\bar{x_1}, \bar{x_2}, \ldots, \bar{x_D}\}}{\arg\min} ||\mathbf{y} - \hat{\mathbf{y}}||_2$$

Geometric Interpretation

Span of ($\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$)

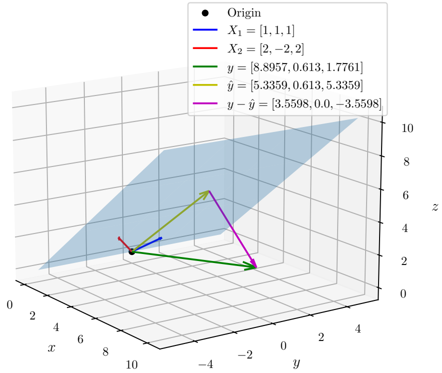


Geometric Interpretation



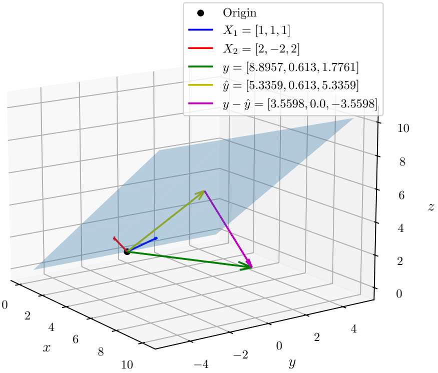
- We seek a $\hat{\mathbf{y}}$ in the span of the columns of $\mathbf{X}$ such that it is closest to $\mathbf{y}$

Geometric Interpretation



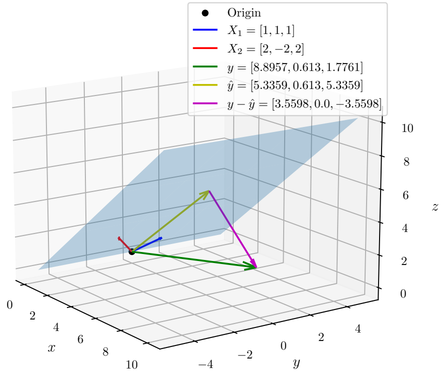
- This happens when $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_j \forall j$ or $\mathbf{x}_j^\top (\mathbf{y} - \hat{\mathbf{y}}) = 0$

Geometric Interpretation



- This happens when $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_j \forall j$ or $\mathbf{x}_j^\top(\mathbf{y} - \hat{\mathbf{y}}) = 0$
- $\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0$

Geometric Interpretation



- This happens when $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_j \forall j$ or $\mathbf{x}_j^\top(\mathbf{y} - \hat{\mathbf{y}}) = 0$
- $\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0$
- $\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\boldsymbol{\theta}$ or $\hat{\boldsymbol{\theta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$

The Problem: Overfitting

- Linear regression can overfit with:

The Problem: Overfitting

- Linear regression can overfit with:
  - Too many features relative to data points

The Problem: Overfitting

- Linear regression can overfit with:
  - Too many features relative to data points
  - Highly correlated features (multicollinearity)

The Problem: Overfitting

- Linear regression can overfit with:
  - Too many features relative to data points
  - Highly correlated features (multicollinearity)
  - Noisy data with complex models

The Problem: Overfitting

- Linear regression can overfit with:
  - Too many features relative to data points
  - Highly correlated features (multicollinearity)
  - Noisy data with complex models
- **Solution**: Add penalty term to control model complexity

The Problem: Overfitting

- Linear regression can overfit with:
  - Too many features relative to data points
  - Highly correlated features (multicollinearity)
  - Noisy data with complex models
- **Solution**: Add penalty term to control model complexity
- This prevents coefficients from becoming too large

Ridge Regression (L2 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \text{MSE} + \lambda \sum_{j=1}^{n} \theta_j^2$$

- $\lambda \geq 0$ is the **regularization parameter**

Ridge Regression (L2 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \text{MSE} + \lambda \sum_{j=1}^{n} \theta_j^2$$

- $\lambda \geq 0$ is the **regularization parameter**
- Larger λ → more regularization → simpler model

Ridge Regression (L2 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \text{MSE} + \lambda \sum_{j=1}^{n} \theta_j^2$$

- $\lambda \geq 0$ is the **regularization parameter**
- Larger λ → more regularization → simpler model
- **Effect**: Shrinks coefficients toward zero

Ridge Regression (L2 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \text{MSE} + \lambda \sum_{j=1}^{n} \theta_j^2$$

- $\lambda \geq 0$ is the **regularization parameter**
- Larger λ → more regularization → simpler model
- **Effect**: Shrinks coefficients toward zero
- **Closed-form solution**: $\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$

Ridge Regression (L2 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \text{MSE} + \lambda \sum_{j=1}^{n} \theta_j^2$$

- $\lambda \geq 0$ is the **regularization parameter**
- Larger λ → more regularization → simpler model
- **Effect**: Shrinks coefficients toward zero
- **Closed-form solution**: $\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$
- Note: $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$ is always invertible for $\lambda > 0$

Lasso Regression (L1 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}(\mathbf{x}^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{n}|\theta_j|}$$

- Uses absolute value penalty instead of squared penalty

Lasso Regression (L1 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\boldsymbol{\theta}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|}$$

- Uses absolute value penalty instead of squared penalty
- **Key Property**: Can set coefficients exactly to zero

Lasso Regression (L1 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_{\boldsymbol{\theta}(\mathbf{x}^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{n}|\theta_j|}$$

- Uses absolute value penalty instead of squared penalty
- **Key Property**: Can set coefficients exactly to zero
- **Automatic Feature Selection**: Eliminates irrelevant features

Lasso Regression (L1 Regularization)

Objective Function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\boldsymbol{\theta}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|}$$

- Uses absolute value penalty instead of squared penalty
- **Key Property**: Can set coefficients exactly to zero
- **Automatic Feature Selection**: Eliminates irrelevant features
- No closed-form solution → requires iterative optimization

Lasso Regression (L1 Regularization)

Objective Function:

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\boldsymbol{\theta}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|}$$

- Uses absolute value penalty instead of squared penalty
- **Key Property**: Can set coefficients exactly to zero
- **Automatic Feature Selection**: Eliminates irrelevant features
- No closed-form solution → requires iterative optimization
- **Use Case**: When you suspect many features are irrelevant

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle
  - Smooth boundary → coefficients shrink smoothly

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle
  - Smooth boundary → coefficients shrink smoothly
  - Rarely sets coefficients exactly to zero

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle
  - Smooth boundary → coefficients shrink smoothly
  - Rarely sets coefficients exactly to zero
- **Lasso (L1)**: Constraint region is a diamond

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle
  - Smooth boundary → coefficients shrink smoothly
  - Rarely sets coefficients exactly to zero
- **Lasso (L1)**: Constraint region is a diamond
  - Sharp corners at axes → coefficients can become exactly zero

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle
  - Smooth boundary → coefficients shrink smoothly
  - Rarely sets coefficients exactly to zero
- **Lasso (L1)**: Constraint region is a diamond
  - Sharp corners at axes → coefficients can become exactly zero
  - Performs automatic feature selection

Ridge vs Lasso: Geometric Intuition

- **Ridge (L2)**: Constraint region is a circle
  - Smooth boundary → coefficients shrink smoothly
  - Rarely sets coefficients exactly to zero
- **Lasso (L1)**: Constraint region is a diamond
  - Sharp corners at axes → coefficients can become exactly zero
  - Performs automatic feature selection
- **Elastic Net**: Combines both penalties

$$J(\boldsymbol{\theta}) = \text{MSE} + \lambda_1 \sum |\theta_j| + \lambda_2 \sum \theta_j^2$$

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization
- λ **very large**: Heavy regularization (underfitting)

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization
- λ **very large**: Heavy regularization (underfitting)
- **Selection Methods**:

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization
- λ **very large**: Heavy regularization (underfitting)
- **Selection Methods**:
  - Cross-validation (most common)

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization
- λ **very large**: Heavy regularization (underfitting)
- **Selection Methods**:
  - Cross-validation (most common)
  - Information criteria (AIC, BIC)

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization
- λ **very large**: Heavy regularization (underfitting)
- **Selection Methods**:
  - Cross-validation (most common)
  - Information criteria (AIC, BIC)
  - Validation curves

Choosing Regularization Parameter λ

- $\lambda = 0$: No regularization (standard linear regression)
- λ **very small**: Minimal regularization
- λ **very large**: Heavy regularization (underfitting)
- **Selection Methods**:
  - Cross-validation (most common)
  - Information criteria (AIC, BIC)
  - Validation curves
- **Critical Insight**: λ controls bias-variance tradeoff

Multi-collinearity

There can be situations where inverse of $X^T X$ is not computable.

Multi-collinearity

There can be situations where inverse of $X^T X$ is not computable.
This condition arises when the $|X^T X| = 0$.

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \tag{1}$$

Multi-collinearity

There can be situations where inverse of $X^T X$ is not computable.
This condition arises when the $|X^T X| = 0$.

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \tag{1}$$

The matrix X is not full rank.

Multi-collinearity

It arises when one or more predictor variables/features in X can be expressed as a linear combination of others

How to tackle it?

- Regularize

Multi-collinearity

It arises when one or more predictor variables/features in X can be expressed as a linear combination of others

How to tackle it?

- Regularize
- Drop variables

Multi-collinearity

It arises when one or more predictor variables/features in X can be expressed as a linear combination of others

How to tackle it?

- Regularize
- Drop variables
- Avoid dummy variable trap

Dummy variables

Say Pollution in Delhi = P

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.
It is denoted as follows {N:0, E:1, W:2, S:3 }

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.
It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?

Dummy variables

Say Pollution in Delhi = P

P = θ_0 + θ_1*#Vehicles + θ_1* *Wind speed* + θ_3 * *Wind Direction*

But, wind direction is a categorical variable.
It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?
Then this implies that S$>$W$>$E$>$N

Dummy Variables

N-1 Variable encoding

	Is it N?	Is it E?	Is it W?
N	1	0	0
E	0	1	0
W	0	0	1
S	0	0	0

Dummy Variables

N Variable encoding

	Is it N?	Is it E?	Is it W?	Is it S?
N	1	0	0	0
E	0	1	0	0
W	0	0	1	0
S	0	0	0	1

Dummy Variables

Which is better N variable encoding or N-1 variable encoding?

Dummy Variables

Which is better N variable encoding or N-1 variable encoding?
The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity.

Dummy Variables

Which is better N variable encoding or N-1 variable encoding?
The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity.
Is it S = 1 - (Is it N + Is it W + Is it E)

Binary Encoding

N	00
E	01
W	10
S	11

Binary Encoding

N	00
E	01
W	10
S	11

W and S are related by one bit.

Binary Encoding

N	00
E	01
W	10
S	11

W and S are related by one bit.
This introduces dependencies between them, and this can cause confusion in classifiers.

Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

Is Female	height
1	...
1	...
1	...
0	...
0	...

Interpreting Dummy Variables

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$

We get $\theta_0 = 5.9$ and $\theta_1 = -0.7$

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$

We get $\theta_0 = 5.9$ and $\theta_1 = -0.7$
θ_0 = Avg height of Male = 5.9
$\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$height_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9
$\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).
θ_1 is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9

Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

height$_i$ = θ_0 + θ_1 * (Is Female) + ϵ_i

We get θ_0 = 5.9 and θ_1 = -0.7
θ_0 = Avg height of Male = 5.9
$\theta_0 + \theta_1$ is chosen based (equal to) on 5, 5.2, 5.4 (for three records).
θ_1 is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9 θ_1 = Avg. female height (5+5.2+5.4)/3 - Avg. male height(5.9)

Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \theta_0 - \theta_1 + \epsilon_i & \text{if } i \text{ th person is male.} \end{cases}$$

Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \theta_0 - \theta_1 + \epsilon_i & \text{if } i \text{ th person is male.} \end{cases}$$

Now, θ_0 can be interpreted as average person height. θ_1 as the amount that female height is above average and male height is below average.

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?
2. When does the normal equation have a unique solution?

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?
2. When does the normal equation have a unique solution?

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?
2. When does the normal equation have a unique solution?
3. How do polynomial features help with non-linear relationships?

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?
2. When does the normal equation have a unique solution?
3. How do polynomial features help with non-linear relationships?

Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?
2. When does the normal equation have a unique solution?
3. How do polynomial features help with non-linear relationships?
4. What are the assumptions behind linear regression?

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between $\mathbf{x}$ and y is linear

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between x and y is linear
- **Independence**: Observations are independent of each other

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between x and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of x

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between x and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of x
- **Normality**: Errors are normally distributed (for inference)

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between $\mathbf{x}$ and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of $\mathbf{x}$
- **Normality**: Errors are normally distributed (for inference)
- **No Multicollinearity**: Features are not highly correlated

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between $\mathbf{x}$ and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of $\mathbf{x}$
- **Normality**: Errors are normally distributed (for inference)
- **No Multicollinearity**: Features are not highly correlated

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between x and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of x
- **Normality**: Errors are normally distributed (for inference)
- **No Multicollinearity**: Features are not highly correlated

Violation Consequences:

- Biased coefficient estimates

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between x and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of x
- **Normality**: Errors are normally distributed (for inference)
- **No Multicollinearity**: Features are not highly correlated

Violation Consequences:

- Biased coefficient estimates
- Invalid confidence intervals

Critical Assumptions of Linear Regression

Before using linear regression, verify these assumptions:

- **Linearity**: Relationship between $\mathbf{x}$ and y is linear
- **Independence**: Observations are independent of each other
- **Homoscedasticity**: Error variance is constant across all values of $\mathbf{x}$
- **Normality**: Errors are normally distributed (for inference)
- **No Multicollinearity**: Features are not highly correlated

Violation Consequences:

- Biased coefficient estimates
- Invalid confidence intervals
- Poor prediction performance

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top \mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix
- **Feature Engineering**: Basis expansion enables non-linear modeling

Key Takeaways

- **Linear Model**: Assumes linear relationship between features and target
- **Least Squares**: Minimizes sum of squared residuals
- **Normal Equation**: Closed-form solution when $\mathbf{X}^\top\mathbf{X}$ is invertible
- **Geometric View**: Projection onto column space of design matrix
- **Feature Engineering**: Basis expansion enables non-linear modeling
- **Foundation**: Building block for more complex models